W1193

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-160928
(43)Date of publication of application : 20.06.1997

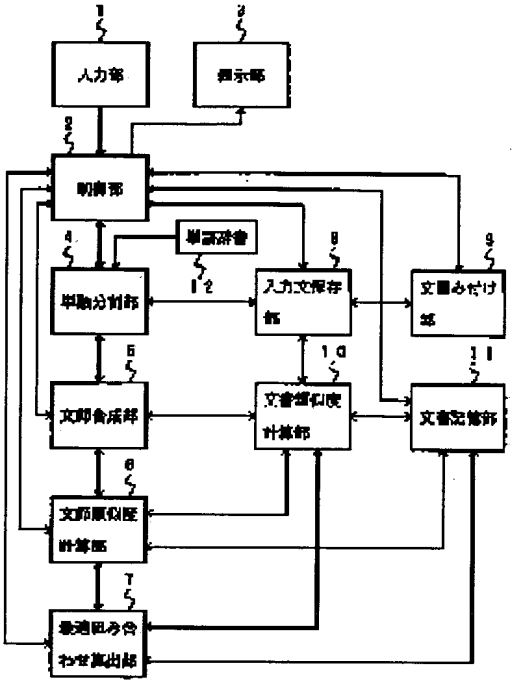
(51)Int.Cl. G06F 17/30

(21)Application number : 07-320544
(22)Date of filing : 08.12.1995

(71)Applicant : TOSHIBA CORP
(72)Inventor : YAJIMA MASATO
KOYAMA NORIKO

(54) DOCUMENT RETRIEVING METHOD AND DEVICE THEREFOR

(57)Abstract:
PROBLEM TO BE SOLVED: To retrieve a document having contents similar to that of the document that a user desires.
SOLUTION: This device is provided with a document similarity calculation part 10 finding the degree of similarity of the whole of the document between the input document to be the retrieval key preserved in an input sentence preservation part 8 and each document (retrieval object document) stored in a document storage part 11. When a retrieval request command is inputted from an input part 1, a sentence is taken out as a retrieval input sentence from the input sentence preservation part 8, a document is taken out as a retrieval object document from the document storage part 11 and the degree of similarity for every paragraph between the retrieval input sentence and each sentence of the retrieval object document is made to be calculated by a paragraph similarity calculation part 6. From the combination result of an optimum combination calculation part 7 based on the calculation result, the degrees of similarity of the whole of each sentence between the retrieval input sentence and each sentence of the retrieval object document are found and are successively added. By performing this operation for all the sentences in the input document, the degree of similarity of the whole of the document between the input document and the retrieval object document is found and a document retrieval is performed based on the similarity.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平9－160928

(43)公開日　平成９年(1997)６月20日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ　15/403	３４０Ｚ ３５０Ｃ

審査請求　未請求　請求項の数６　ＯＬ　（全 14 頁）

(21)出願番号　特願平7－320544

(22)出願日　平成７年(1995)12月８日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　矢島　真人
神奈川県川崎市幸区柳町70番地　株式会社東芝柳町工場内

(72)発明者　小山　紀子
神奈川県川崎市幸区柳町70番地　株式会社東芝柳町工場内

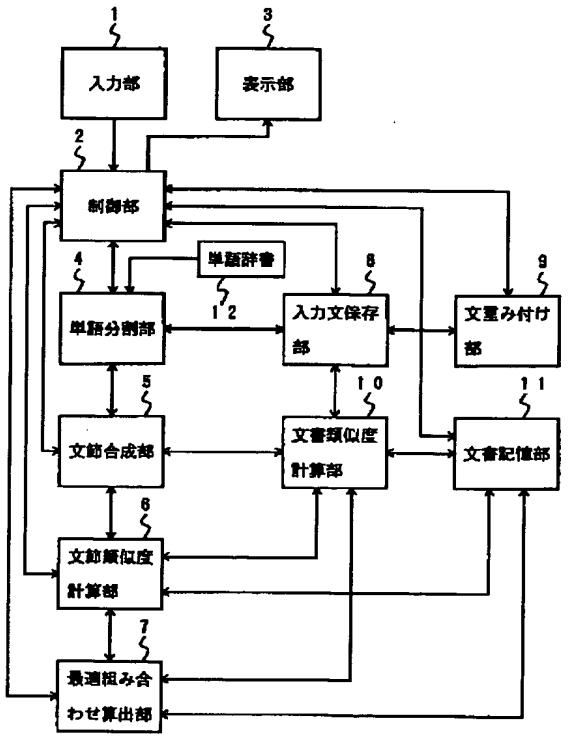
(74)代理人　弁理士　鈴江　武彦

(54)【発明の名称】　文書検索方法及び装置

(57)【要約】

【課題】ユーザが望む文書と内容が類似した文書を検索できるようにする。

【解決手段】入力文保存部８に保存された検索キーとなる入力文書と、文書記憶部１１に記憶されている各文書（検索対象文書）との間の文書全体の類似度を求める文書類似度計算部１０を設け、入力部１から検索要求コマンドが入力された場合には、入力文保存部８から１文を検索入力文として取り出すと共に、文書記憶部１１から１文書を検索対象文書として取り出し、検索入力文と検索対象文書の各文との間の文節ごとの類似度を文節類似度計算部６により計算させ、その計算結果に基づく最適組み合わせ算出部７の組み合わせ結果から、検索入力文と検索対象文書の各文との間のそれぞれ文全体の類似度を求めて順次加算する操作を、入力文書のすべての文について行うことで、入力文書と検索対象文書との間の文書全体の類似度を求め、その類似度をもとに文書検索を行う。



【特許請求の範囲】

【請求項１】　指定された文書中のすべての文を検索入力文として、検索対象として用意されている文書中の各文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算して、その文全体の類似度を求め、

その文全体の類似度を順次加算した値を前記指定文書と検索対象文書との間の文書全体の類似度として求める操作を、検索対象となるすべての文書について順次行うことにより、

前記求めた文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索方法。

【請求項２】　複数の文書を記憶しておくための文書記憶手段と、

検索入力文と検索対象文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算する文構成単位類似度計算手段と、

この文構成単位類似度計算手段の計算結果をもとに、前記検索入力文と検索対象文との間で最も類似している文構成単位の組み合わせを求める最適組み合わせ算出手段と、

指定された文書から１文を検索入力文として取り出すと共に、前記文書記憶手段から１文書を検索対象文書として取り出して、前記検索入力文と前記検索対象文書の各文との間の文構成単位の類似度を、前記文構成単位類似度計算手段により計算させ、その計算結果に基づく前記最適組み合わせ算出手段の組み合わせ結果から、前記検索入力文と前記検索対象文書の各文との間のそれぞれ文全体の類似度を求めて順次加算する操作を、前記指定文書のすべての文について行うことで、前記指定文書と検索対象文書との間の文書全体の類似度を求める文書類似度計算手段とを具備し、

前記指定文書との文書全体の類似度を前記文書記憶手段内の検索対象となる複数の文書のそれぞれについて求めることにより、当該文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索装置。

【請求項３】　単語辞書内の各単語ごとにその重要度を付与しておき、

指定された文書で使用されている各単語の重要度を前記単語辞書から求めることで、その各単語の重要度に応じて前記指定文書中の各文に重み付けを行い、

前記指定文書中のすべての文を検索入力文として、検索対象として用意されている文書中の各文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算して、その文全体の類似度を求め、

その文全体の類似度と当該文に付された重みとの乗算値を順次加算した値を前記指定文書と検索対象の文書との間の文書全体の類似度として求める操作を、検索対象となるすべての文書について行うことにより、

前記求めた文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索方法。

【請求項４】　重要度を示す情報が付された複数の単語を登録しておくための単語辞書と、

この単語辞書の示す、指定された文書で使用されている各単語の重要度に応じて前記指定文書中の各文に重み付けを行う文重み付け手段と、

複数の文書を記憶しておくための文書記憶手段と、

検索入力文と検索対象文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算する文構成単位類似度計算手段と、

この文構成単位類似度計算手段の計算結果をもとに、前記検索入力文と検索対象文との間で最も類似している文構成単位の組み合わせを求める最適組み合わせ算出手段と、

前記指定文書から１文を検索入力文として取り出すと共に、前記文書記憶手段から１文書を検索対象文書として取り出して、前記検索入力文と前記検索対象文書の各文との間の文構成単位の類似度を、前記文構成単位類似度計算手段により計算させ、その計算結果に基づく前記最適組み合わせ算出手段の組み合わせ結果から、前記検索入力文と前記検索対象文書の各文との間のそれぞれ文全体の類似度を求め、その文全体の類似度と前記文重み付け手段により当該文に付された重みとの乗算値を順次加算する操作を、前記指定文書のすべての文について行うことで、前記指定文書と検索対象文書との間の文書全体の類似度を求める文書類似度計算手段とを具備し、

前記指定文書との文書全体の類似度を前記文書記憶手段内の検索対象となる複数の文書のそれぞれについて求めることにより、当該文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索装置。

【請求項５】　指定された文書で使用されている各単語の頻度を求めることで、その各単語の頻度に応じて前記指定文書中の各文に重み付けを行い、

前記指定文書中のすべての文を検索入力文として、検索対象として用意されている文書中の各文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算して、その文全体の類似度を求め、

その文全体の類似度と当該文に付された重みとの乗算値を順次加算した値を前記指定文書と検索対象の文書との間の文書全体の類似度として求める操作を、検索対象となるすべての文書について行うことにより、

前記求めた文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索方法。

【請求項６】　指定された文書で使用されている各単語の頻度を求めることで、その各単語の頻度に応じて前記

指定文書中の各文に重み付けを行う文重み付け手段と、
複数の文書を記憶しておくための文書記憶手段と、
検索入力文と検索対象文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算する文構成単位類似度計算手段と、
この文構成単位類似度計算手段の計算結果をもとに、前記検索入力文と検索対象文との間で最も類似している文構成単位の組み合わせを求める最適組み合わせ算出手段と、
前記指定文書から１文を検索入力文として取り出すと共に、前記文書記憶手段から１文書を検索対象文書として取り出して、前記検索入力文と前記検索対象文書の各文との間の文構成単位の類似度を、前記文構成単位類似度計算手段により計算させ、その計算結果に基づく前記最適組み合わせ算出手段の組み合わせ結果から、前記検索入力文と前記検索対象文書の各文との間のそれぞれ文全体の類似度を求め、その文全体の類似度と前記文重み付け手段により当該文に付された重みとの乗算値を順次加算する操作を、前記指定文書のすべての文について行うことで、前記指定文書と検索対象文書との間の文書全体の類似度を求める文書類似度計算手段とを具備し、
前記指定文書との文書全体の類似度を前記文書記憶手段内の検索対象となる複数の文書のそれぞれについて求めることにより、当該文書全体の類似度をもとに、前記検索対象となる複数の文書の中から前記指定文書と類似した文書を検索することを特徴とする文書検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、検索の対象となる大量の文書中からユーザ指定の文書と類似している文書を検索するのに好適な文書検索方法及び装置に関する。

【０００２】

【従来の技術】日本語ワードプロセッサや光学的文字読取装置（ＯＣＲ）などの文書入力装置の普及に伴い、従来紙やマイクロフィルムという形態で保存していた文書を、電子化して磁気ディスクや光磁気ディスクなどの外部記憶装置に保存するようになってきた。また大量の電子化された文書データから、いかにユーザが要求する文書を検索するかというテキスト検索技術も開発され発展しつつある。

【０００３】従来の検索技術では、ユーザが文書を検索するために検索キーとなる文字列や単語などを入力して、検索キーが含まれる文書を検索するという方式をとることが多く、検索キーである文字列や単語が検索文書内でどのように使われているかを問わないので、ユーザが望む文書以外に、意味のない文書を検索してしまうという問題があった。

【０００４】これに対して、特願平５－２１２５１５号にて提案されているように、文を一定の文構成単位、例えば文節単位に分割して文節ごとに類似度を計算し、最も類似している文節の組み合わせを求めるという検索方式がある。この方式では、入力した文（ユーザ指定の文）に対して文全体が類似している文を検索できるという特長がある。これによると、単なる文字列や単語による検索と異なり１文を検索キーとするため、ユーザが期待する文章表現と類似した文を含む文書を検索することができる。

【０００５】しかし、この方式では、例えばユーザが特定の文書と内容が類似した文書を検索したいと考えた場合、当該特定の文書内の各文に対して類似した文を含む文書を検索することはできるが、文書全体が類似した文書とは限らないため、やはり意味のない文書を検索してしまうという可能性を免れなかった。

【０００６】

【発明が解決しようとする課題】上記したように従来の文書検索方式では、ユーザ指定の１文と類似した文を含む文書を検索することはできるものの、特定の文書に対して類似した文書を検索することはできなかった。

【０００７】本発明は上記事情を考慮してなされたものでその目的は、特定の文書（ユーザの入力した文書あるいはユーザの指定した文書）内の各文と類似した文を含む文書を検索するだけでなく、文書全体がどれだけ類似しているかを総合的に判断することで、当該特定の文書と内容が類似した文書を検索することができる文書検索方法及び装置を提供することにある。

【０００８】

【課題を解決するための手段】本発明の第１の観点に係る文書検索方法は、指定文書（ユーザが入力した文書あるいはユーザが指定した文書）の中のすべての文を検索入力文として、検索対象として用意されている文書中の各文との間で一定の文構成単位ごとに比較することで当該文構成単位の類似度を計算して、その文全体の類似度を求め、その文全体の類似度を順次加算した値を上記指定文書と検索対象文書との間の文書全体の類似度として求める操作を、検索対象となるすべての文書について行うことにより、この文書全体の類似度をもとに検索対象となる複数の文書の中から上記指定文書と類似した文書を検索することを特徴とする。

【０００９】本発明の第２の観点に係る文書検索方法は、単語辞書内の各単語ごとに重要度を付与しておき、指定文書で使用されている各単語の重要度を上記単語辞書から求めることで、その各単語の重要度に応じて指定文書中の各文に重み付けを行い、上記第１の観点に係る文書検索方法において、文全体の類似度を順次加算した値を文書全体の類似度とするのに代えて、文全体の類似度と当該文に付された重みとの乗算値を順次加算した値を文書全体の類似度とするようにしたことを特徴とする。

【００１０】本発明の第３の観点に係る文書検索方法は、指定文書で使用されている各単語の頻度を求めるこ

とで、その各単語の頻度に応じて前記指定文書中の各文に重み付けを行い、上記第１の観点に係る文書検索方法において、文全体の類似度を順次加算した値を文書全体の類似度とするのに代えて、文全体の類似度と当該文に付された重みとの乗算値を順次加算した値を文書全体の類似度とするようにしたことを特徴とする。

【００１１】本発明においては、指定文書中のすべての文を検索入力文として、検索対象文書中の各文に対して１文全体の類似度を求めて順次加算した値を文書全体の類似度とすることで、指定文書と検索対象文書とが文書全体としてどれだけ類似しているかを、当該文書全体の類似度で表すことができるため、この文書全体の類似度をすべての検索対象文書について求めることにより、この文書全体の類似度をもとにすべての検索対象文書の中から指定文書と類似した文書を検索することが可能となる。

【００１２】また本発明においては、指定文書で使用されている各単語の重要度または使用頻度に応じて当該指定文書中の各文に重み付けを行い、その各文ごとの重みをその文（検索対象文）と検索入力文との文全体の類似度に反映させることで、すべての検索対象文書の中から指定文書と類似した文書であって且つ指定文書中で使用している重要度の高い単語を多数含む、あるいは指定文書中で使用している単語の使用数の多い文書を検索することが可能となる。

【００１３】このように本発明によれば、ユーザ指定の文書と文書全体が類似した文書を検索することができるために、ユーザが特定の文書と類似した文書を検索したい場合に、従来のように当該特定の文書内で使われている単語やキーワードを選び出して検索したり、重要な文を選び出して類似した文を含む文書を検索したりする手間は一切いらなくなり、単に手元にある文書をそのまま指定文書として入力したり、記憶手段内の文書群の中から所望の文書を指定文書として呼び出すだけで、ユーザが意図する文書を検索することが可能となる。

【００１４】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。図１は本発明の一実施形態に係る文書検索装置の全体構成を示すブロック図である。

【００１５】図１の文書検索装置は例えばパーソナルコンピュータ等の情報処理装置により実現されるもので、入力部１、制御部２、表示部３、単語分割部４、文節合成部５、文節類似度計算部６、最適組み合わせ算出部７、入力文保存部８、文重み付け部９、文書類似度計算部１０、文書記憶部１１及び単語辞書１２から構成される。

【００１６】入力部１は、文字列や種々の制御コード（コマンド等）を入力するためのもので、例えばキーボードまたはＯＣＲ（光学的文字読取装置）からなる入力装置である。

【００１７】制御部２は装置全体を司るもので、例えば中央処理ユニット（ＣＰＵ）である。ここで制御部２は、文書検索処理を実行するための装置内各部（単語分割部４、文節合成部５、文節類似度計算部６、最適組み合わせ算出部７、文重み付け部９、文書類似度計算部１０）の制御を行う。

【００１８】表示部３は、入力文の文字列や検索対象の文書などを表示するためのもので、例えばＣＲＴディスプレイ装置または液晶表示装置（等のフラットパネルディスプレイ装置）である。

【００１９】入力文保存部８は、検索のキーとなる入力文書（指定文書）を一時的に保存するためのもので、例えばメモリなどからなる内部記憶装置により実現される。文書記憶部１１は、検索対象となる複数の文書を記憶しておくためのもので、例えばハードディスク装置や光ディスク装置などからなる大容量の外部記憶装置により実現される。

【００２０】単語辞書１２は、多数の単語を、その単語（を構成する文字列）の表記（見出し）、品詞等と対応付けで登録しておくもので、ＲＯＭ等の不揮発性記憶装置に置かれる。この単語辞書１２内の各単語には、その単語の重要度（を示す情報）が付されている。本実施形態において、単語辞書１２内の重要度を含む登録情報は固定であるものとする。このような状態で、上記単語の重要度を変更可能とするためには、例えば単語辞書１２に優先して使用されるユーザ辞書を設け、このユーザ辞書に所望の単語を登録して、その単語の重要度を任意に設定可能とすればよい。

【００２１】単語分割部４、文節合成部５、文節類似度計算部６、最適組み合わせ算出部７、文重み付け部９及び文書類似度計算部１０は、文書検索処理に必要な機能要素であり、それぞれ固有の処理ルーチン（と当該処理ルーチンを実行するＣＰＵ）により実現されるものである。

【００２２】単語分割部４は、入力部１から検索のキーとなる文書を構成する文をなす文字列が入力された場合に、その文字列からなる入力文を検索入力文として単語単位に分割するものである。ここで文とは、句点で区切られる文字列、箇条書きの文字列等、文書を構成する独立の文字列をいう。

【００２３】文節合成部５は、単語分割部４により分割された単語からなる文節を合成するものである。文節類似度計算部６は、文節合成部５により得られた検索入力文の文節と、文書記憶部１１に記憶されている文書（検索対象文書）の１文（検索対象文）の文節との類似度計算を行うものである。

【００２４】最適組み合わせ算出部７は、文節類似度計算部６での類似度計算結果に基づいて、文全体として最も類似している文節の組み合わせを求めるものである。文重み付け部９は、入力文保存部８に保存された入力文

書の各文について、その文書中で使用されている単語の重要度または頻度（使用頻度）に応じて重み付けを行うものである。

【００２５】文書類似度計算部１０は、入力文保存部８に保存された入力文書のすべての文（検索入力文）について、文書記憶部１１に記憶されている文書（検索対象文書）の各文（検索対象文）との類似度を計算して、順次加算し、入力文書（指定文書）と検索対象文書との文書全体としての類似度を算出するものである。

【００２６】次に、図１の構成の動作について、その概要を説明する。まず入力部１から、検索のキーとなる文書（入力文書）を構成する文（検索入力文）をなす文字列が入力されると、当該文字列（入力文）は制御部２を介して単語分割部４に送られる。単語分割部４は、単語辞書１２を参照して、この入力文を単語単位に分割する。このとき単語分割部４は、分割した単語単位に、品詞、及び活用形、時制などの（形態素情報と呼ばれる）単語属性情報、及び重要度を付加する。

【００２７】単語分割部４が入力文を単語単位に分割した結果は、入力文保存部８に保存されると共に、文節合成部５に送られる。文節合成部５は、単語分割部４により分割された単語から文節を合成する。

【００２８】文節合成部５が文節合成した結果は、入力文保存部８に送られ、先に保存された単語分割結果に対応付けて保存される。以上の処理は、入力部１から順次入力される各文について繰り返され、単語ごとに単語属性情報及び重要度が付加され、さらに文節合成された状態で、それぞれ検索のキーとなる入力文書（指定文書）の検索入力文として入力文保存部８に保存される。

【００２９】さて、入力文保存部８に検索入力文の群からなる入力文書が保存されている状態で、入力部１から、検索要求のコマンドが入力されると、まず制御部２により文重み付け部９が起動される。すると文重み付け部９は、入力文保存部８に保存された各検索入力文それぞれの重みを、対応する入力文書中で使用されている単語の重要度または頻度をもとに計算し、文書類似度計算部１０に送る。

【００３０】文書類似度計算部１０は、文書記憶部１１から検索対象となる文書を先頭から１文書取り出し、さらに、その１文書の先頭から１文（検索対象文）を取り出して文節類似度計算部６に送る。また文書類似度計算部１０は、入力文保存部８に保存された入力文書中の文（検索入力文）を先頭から１文取り出して文節類似度計算部６に送る。

【００３１】文節類似度計算部６は、文書類似度計算部１０から送られた入力文の（文書構成単位としての）文節と、検索対象文の（文書構成単位としての）文節とを比較して類似度を計算する動作を、（入力文の文節数）×（検索対象文の文節数）の総当たりで行う。最適組み合わせ算出部７は、この文節類似度計算部６での各文節ごとの類似度計算結果に基づいて、文全体として最も類似している文節の組み合わせを求める。

【００３２】文書類似度計算部１０は、最適組み合わせ算出部７での文節組み合わせ結果から文全体の類似度を計算し、その計算結果（類似度）と、文重み付け部９により算出され（て当該文重み付け部９から送られ）た対応する入力文の重みとの積（乗算値）を求め、文書全体の類似度（初期値は０）に加算する。

【００３３】文書類似度計算部１０は、文書記憶部１１から取り出した文書（検索対象文書）に未処理の文（検索対象文）が残っているか否かを調べ、残っているならば次の文を取り出して、上記入力文（検索入力文）との類似度を計算し、その計算結果と対応する入力文の重みとの積を求めて、文書全体の類似度に加算する動作を繰り返す。

【００３４】一方、文書記憶部１１から取り出した文書に未処理の文が残っていないならば、文書類似度計算部１０は、入力文保存部８に次の入力文（未処理の入力文）があるか否かを調べ、あるならば文書記憶部１１から取り出した文書の先頭から１文を取り出して、当該次の入力文との類似度を計算し、その計算結果と対応する入力文の重みとの積を求めて、文書全体の類似度に加算する動作を繰り返す。

【００３５】やがて、入力文保存部８に次の入力文がなくなると、文書類似度計算部１０は、入力文保存部８に保存されている文書（入力文書）と文書記憶部１１から取り出した文書（検索対象文書）との間の文書全体の類似度が求まったものと判断する。

【００３６】この場合、文書類似度計算部１０は、求まった文書全体の類似度から、該当する検索対象文書が入力文書に類似しているか否かを判断し、類似していると判断できたならば、当該検索対象文書、即ち文書記憶部１１から取り出した文書を、入力文書に類似した文書であるとして表示部３に表示する。

【００３７】次に文書類似度計算部１０は、文書記憶部１１に未処理の文書（検索対象文書）が残っているか否かを調べ、残っているならば次の文書を取り出す。そして文書類似度計算部１０は、入力文保存部８に保存された入力文書中の文（検索入力文）を先頭から１文取り出すと共に、上記文書記憶部１１から取り出した検索対象文書の先頭から１文（検索対象文）を取り出して、入力文と検索対象文との類似度を求め、その類似度と対応する入力文の重みとの積を求めて、その検索対象文書（と入力文書）との間の文書全体の類似度に加算する動作を繰り返す。

【００３８】やがて、文書記憶部１１内に未処理の文書がなくなると、上記した一連の検索処理は終了する。以上が、図１の構成の動作の概要である。

【００３９】次に、図１の構成の動作の詳細について、図２乃至図４のフローチャートを参照して説明する。ま

ず制御部２は、入力部１から検索キーとなる文書の１文（検索入力文）をなす文字列が入力されたと判断すると（ステップＳ１）、当該入力文（入力文字列）を入力し（ステップＳ２）、表示部３の表示画面に表示すると共に、単語分割部４に送る。

【００４０】ここでは、入力文として、図５（ａ）において符号ａ１で示すような「Ａ社は、低価格のノート型ＥＷＳを開発した。」という文字列が入力され、単語分割部４に送られたとする。

【００４１】単語分割部４は、制御部２から送られた入力文（検索入力文）ａ１の文字列を単語辞書１２の登録情報を用いて解析することで単語単位に分割する（ステップＳ３）。

【００４２】ここでは単語分割部４は、図５（ａ）の入力文ａ１を、図６（ａ）に示すように、「Ａ社」「は」「低価格」「の」「ノート型」「ＥＷＳ」「を」「開発する」のように単語に分割し、それぞれの品詞、活用形及び時制などの単語属性の情報、及び（単語辞書１２に登録されている）当該単語の重要度を付加する。この図６（ａ）に示したような単語分割部４による単語分割結果は、入力文保存部８に保存されると共に文節合成部５に渡される。

【００４３】単語分割部４による単語分割処理が終了すると、制御部２により文節合成部５が起動される。文節合成部５は、単語分割部４により得られた単語を合成して、文節を生成する（ステップＳ４）。

【００４４】ここでは、図６（ａ）に示す単語分割処理結果から、図６（ｂ）のように、「Ａ社は」「低価格の」「ノート型」「ＥＷＳを」「開発する」の５文節が生成される。

【００４５】文節合成部５による文節単位への合成処理が終了すると、図６（ｂ）に示したような当該文節合成の結果が、図６（ａ）に示したような単語分割結果に対応付けて、制御部２により入力文保存部８に保存される（ステップＳ５）。

【００４６】以上のステップＳ１からＳ５までの処理を繰り返すことで、入力文保存部８には入力部１から入力された入力文（検索入力文）が図６（ａ），（ｂ）に示したような形式で順次保存される。

【００４７】図５（ａ）は、このような入力部１から順次文を入力してできた入力文書を示すもので、上記した入力文ａ１の他に、入力文ａ２，ａ３から構成されるものとする。

【００４８】次に、入力文保存部８に入力文書が保存されている状態で、入力部１から検索要求のコマンドが入力されたものとする。制御部２は、入力部１から検索要求のコマンドが入力されたことを検出すると（ステップＳ１）、文重み付け部９を起動する。すると文重み付け部９は、入力文保存部８に保存されている入力文書の各文（検索入力文）それぞれの重みを、その文書中で使用されている単語の重要度または頻度をもとに算出するする重み付け処理を行う（ステップＳ６）。

【００４９】以下、この文重み付け部９によるステップＳ６の重み付け処理の詳細を、入力文保存部８に保存されている文書中で使用されている単語の重要度から各文に重み付けをする場合を例に、図４（ａ）のフローチャートを参照して説明する。

【００５０】文重み付け部９は、入力文保存部８に保存されている入力文書の各文ごとに、その文で使われている単語の重要度を加算して、各文ごとの重みを求める（ステップＳ２１）。ここで各文で使われている単語の重要度は、前記ステップＳ３で単語分割部４が単語単位の分割を行った時点で、図６（ａ）の例のように単語ごとに付加されて、入力文保存部８に保存されたものである。

【００５１】図５（ａ）に示した入力文ａ１～ａ３からなる文書が入力文保存部８に保存されている例では、先頭の「Ａ社は、低価格のノート型ＥＷＳを開発した。」という入力文ａ１中の各単語の重要度は、図６（ａ）から明らかなように、「Ａ社」＝１、「ノート型」＝２、「ＥＷＳ」＝１なので（他の単語の重要度はすべて０）、その重要度を加算した結果は１＋２＋１＝４となり、これが先頭の入力文ａ１の重みとされる。

【００５２】同様にして、後続の入力文ａ２，ａ３についても、その入力文ａ２，ａ３中の各単語の重要度を加算する処理が行われて、その加算結果が入力文ａ２，ａ３の重みとされる。

【００５３】図７（ａ）は、このようにして求められた（入力文保存部８に保存されている文書中の）各入力文ａ１～ａ３の重みの一例を示す。次に、上記ステップＳ６の処理の詳細を、入力文保存部８に保存されている文書に使われている単語の頻度から各文に重み付けをする場合を例に、図４（ｂ）のフローチャートを参照して説明する。

【００５４】まず文重み付け部９は、入力文保存部８に保存されている文書（ここでは、図５（ａ）に示すように文ａ１～ａ３からなる文書）で使用されている各単語の頻度（使用頻度）を調べる（ステップＳ３１）。

【００５５】ここでは、各単語の頻度を上記した単語の重要度と同様に扱うようにしており、１文書中で１回した使用されない単語（頻度１の単語）については対象外とする。したがって、図５（ａ）の文書の例では、各単語の頻度として、「Ａ社」＝２、「ノート型」＝２、「ＥＷＳ」＝３、「Ｂ社」＝２が求められる。

【００５６】次に文重み付け部９は、入力文保存部８に保存されている文書の各文ごとに、その文で使われている各単語の頻度を加算して、各文ごとの重みを求める（ステップ３２）。

【００５７】図５（ａ）に示した入力文ａ１～ａ３からなる文書が入力文保存部８に保存されている例では、先

頭の「Ａ社は、低価格のノート型ＥＷＳを開発した。」という入力文（第１文）ａ１の重みは、当該入力文ａ１中の単語の頻度が「Ａ社」＝２、「ノート型」＝２、「ＥＷＳ」＝３であることから、２＋２＋３＝７となる。

【００５８】同様に、「ノート型ＥＷＳは、すでにＢ社が今年５月に発売した。」という入力文（第２文）ａ２の重みは、当該入力文ａ２中の単語の頻度が「ノート型」＝２、「ＥＷＳ」＝３、「Ｂ社」＝２であることから、２＋３＋２＝７となり、「Ａ社のＥＷＳは、Ｂ社よりも１０％程度価格が下がった。」という入力文（第３文）ａ３の重みは、当該入力文ａ３中の単語の頻度が、「Ａ社」＝２、「ＥＷＳ」＝３、「Ｂ社」＝２の頻度から２＋３＋２＝７となる以上が文重み付け部９によるステップＳ６の処理の詳細である。

【００５９】文重み付け部９によるステップ６の処理が行われて、入力文保存部８に保存されている文書の各文ごとの重みが求められると、制御部２により文書類似度計算部１０が起動される。

【００６０】文書類似度計算部１０は、文書記憶部１１から検索対象となる文書を先頭から順に１文書取り出す（ステップＳ７）。ここでは、文書記憶部１１から、図５（ｂ）に示すような、文ｂ１，ｂ２からなる先頭の文書が取り出されたものとする。但し、文書記憶部１１に実際に格納されている文書（中の各文）の形式は、入力文保存部８に保存されている文書（中の各文）と同様であり、図５（ｂ）に示したような単なる文字列（文字コード列）の形ではない。即ち文書記憶部１１には、各文書中の各文が、あらかじめ単語分割部４の処理（または単語分割部４と同等の機能）により分割された（単語属性情報及び重要度が付加された）単語列と、文節合成部５の処理（または文節合成部５と同等の機能）により生成された文節列の形で格納されている。

【００６１】文書類似度計算部１０は、ステップＳ７で文書記憶部１１から検索対象となる１文書を取り出すと、今度は入力文保存部８から先頭の１文（検索入力文）を取り出すと共に（ステップＳ８）、ステップＳ７で文書記憶部１１から取り出した検索対象文書の先頭から１文を検索対象文として取り出し（ステップＳ９）、その取り出した検索入力文と検索対象文の文節を文節類似度計算部６に渡す。

【００６２】文節類似度計算部６は、文書類似度計算部１０がステップＳ８で取り出した検索入力文の文節とステップＳ９で取り出した文書記憶部１１中の文書（検索対象文書）の１文（検索対象文）の文節とを当該文書類似度計算部１０から受け取ると、その検索入力文の１文節と検索対象文の１文節とを比較して、その文節間の類似度を計算する（ステップＳ１０）。この文節類似度計算部６での類似度計算は、（検索入力文の文節数）×（検索対象文の文節数）の総当たりで行われ、すべての文節について繰り返される（ステップＳ１１）。

【００６３】最適組み合わせ算出部７は、文節類似度計算部６による検索入力文と検索対象文との文節間の類似度計算の結果に基づいて、当該検索入力文と検索対象文との間で文全体として最も類似している文節の組み合わせ（最適組み合わせ）を求める（ステップＳ１２）。

【００６４】ここでは、図５（ａ）に示す文書中の入力文（検索入力文）ａ１と、図５（ｂ）に示す文書中の第１文（検索対象文）ｂ１の文節間の類似度計算が行われ、その文ａ１，ｂ１間で文全体として最も類似している文節の組み合わせとして、「Ａ社は」と「Ａ社は」、「低価格の」と「低価格の」、及び「ＥＷＳを」と「ＥＷＳを」の３文節が求められたものとする。この最適組み合わせ算出部７での最適組み合わせ算出結果は、文書類似度計算部１０に渡される。

【００６５】文書類似度計算部１０は、最適組み合わせ算出部７から最適組み合わせ算出結果を受け取ると、その結果をもとに、例えば検索入力文中の文節数に対する最も類似している文節の組み合わせ数の割合を求めて文全体の類似度とし、それにステップＳ６で文重み付け部９により求められている対応する検索入力文の重みを乗じた値（文全体の類似度と検索対象文書の重みとの積）を算出する（ステップＳ１３）。

【００６６】ここでは、最適組み合わせ算出部７の処理より、検索入力文ａ１の５文節中「Ａ社は」「低価格の」「ＥＷＳを」の３文節が検索対象文ｂ１と一致していることが求められているため、その割合（最も類似している文節の組み合わせ数の割合）３／５＝０．６が、検索入力文ａ１と検索対象文ｂ１との間の文全体の類似度として計算される。また、文重み付け部９で求められた検索入力文ａ１の重みは、例えば単語の重要度を用いた重み付けの例では、４である。この場合、ステップＳ１３で求められる、検索入力文ａ１と検索対象文ｂ１の間の類似度と検索入力文ａ１の重みとの積は、０．６×４＝２．４となる。

【００６７】次に文書類似度計算部１０は、ステップＳ１３で求めた値（文全体の類似度と重みとの乗算値）を、文書全体の類似度（ここでは、入力文保存部８に保存されている文書とステップＳ７で文書記憶部１１から取り出された検索対象文書との間の文書全体の類似度）に加算する（ステップＳ１４）。この文書全体の類似度の初期値は０であり、したがって１回目の加算結果は、ステップＳ１３で求めた値に一致する。

【００６８】文書類似度計算部１０はステップＳ１４を実行すると、ステップＳ７で取り出した検索対象文書の中に未処理の文、即ちステップＳ９で未だ取り出されていない文が残っているか否かを調べる（ステップＳ１５）。

【００６９】もし、未処理の文があるならば、文書類似度計算部１０はステップＳ９に戻り、ステップＳ７で取

り出した検索対象文書の中から次の１文を検索対象文として取り出す。ここで取り出される次の文は、図５（ｂ）に示す文書中の第２文（最後の検索対象文）ｂ２である。

【００７０】さて、検索入力文ａ１と検索対象文ｂ２との間では、「ノート型の」と「ノート型の」、「ＥＷＳ」と「ＥＷＳ」の２文節が対応している。このため、上記ステップＳ９に続くステップＳ１０～Ｓ１２を経た後、検索入力文ａ１中の文節数に対する最も類似している文節の組み合わせ数の割合２／５＝０．４が文全体の類似度として求められ、これに検索入力文ａ１の重み４を乗じた値０．４×４＝１．６が（その時点における）文書全体の類似度に加算される（ステップＳ１３，Ｓ１４）。ここでは、加算前の文書全体の類似度が２．４であることから、新たな文書全体の類似度は２．４＋１．６＝４．０となる。

【００７１】ステップＳ１４が実行されると、前記したように、ステップＳ７で取り出した検索対象文書の中に未処理の文が残っているか否かが調べられる（ステップＳ１５）。

【００７２】この例のように未処理の文がないならば、文書類似度計算部１０は、入力文保存部８に保存されている文書の中に未処理の文、即ちステップＳ８で取り出した検索入力文が残っているか否かを調べる（ステップＳ１６）。

【００７３】もし、未処理の文があるならば、文書類似度計算部１０はステップＳ８に戻り、入力文保存部８に保存されている文書の中から次の１文を検索入力文として取り出す。ここで取り出される次の文は、図５（ａ）に示す文書中の第２文（検索入力文）ａ２である。

【００７４】文書類似度計算部１０は、ステップＳ８で次の検索入力文ａ２を取り出すと、前記ステップＳ９以降の処理を繰り返す。これにより、図５（ａ）中の検索入力文ａ２に対して図５（ｂ）中の各検索対象文ｂ１，ｂ２の類似度が順次計算され、その都度当該類似度に検索入力文ａ２の重み（ここでは図７（ａ）に示すように３）を乗じた値が求められて、文書全体の類似度に加算される。

【００７５】以下、同様にして、入力文保存部８に保存されている文書中の第３文（最後の検索入力文）ａ３がステップＳ８で取り出され、ステップＳ９以降の処理が繰り返される。すると、図５（ａ）中の検索入力文ａ３に対して図５（ｂ）中の各検索対象文ｂ１，ｂ２の類似度が順次計算され、その都度当該類似度に検索入力文ａ３の重み（ここでは図７（ａ）に示すように２）を乗じた値が求められて、文書全体の類似度に加算される。

【００７６】この例では、検索入力文ａ１～ａ３と検索対象文ｂ１，ｂ２との間の文全体の類似度（検索入力文中の文節数に対する、検索入力文と検索対象文との間で最も類似している文節の組み合わせ数の割合）は図７（ｂ）のようになる。

【００７７】そして、検索入力文と検索対象文との間の文全体の類似度が求められる都度、上記ステップＳ１３，Ｓ１４にて、その類似度と当該検索入力文の重みとの積の値が、文書全体の類似度に加算されていく。

【００７８】したがって、図７（ｂ）に示したような検索入力文ａ１～ａ３と検索対象文ｂ１，ｂ２との間の文全体の各類似度に、対応する検索入力文の重み（図７（ａ）参照）を乗じ、その乗算値（類似度と重みとの積）を順次加算し終えた段階では、第５（ａ）に示した入力文書と図５（ｂ）に示した検索対象文書との間の文書全体の類似度は、

０．６×４＋０．４×４＋０．２５×３＋０．３７５×３＋０，１４×２＋０．５７×２＝７．２９５

のように求められる。

【００７９】この段階では、ステップＳ７で取り出した検索対象文書の中に未処理の文はなく（ステップＳ１５）、また入力文保存部８に保存された文書の中にも未処理の文はないことから（ステップＳ１６）、文書類似度計算部１０はステップＳ１７に進む。

【００８０】文書類似度計算部１０は、ステップ１７において、その時点において求められている文書全体の類似度（ここでは、７．２９５）から、入力文保存部８に保存された入力文書と文書記憶部１１から取り出された検索対象文書とが類似しているか否かを判断する。ここでの判断基準は、設定された基準値（閾値）を越える類似度を持つか否かであり、この基準値は、ユーザの要求する検索精度に応じて適宜変更設定可能である。例えば、最低の検索精度でよい場合には、上記基準値は０に設定される。この場合、文書全体の類似度として０以外の値が求められているならば、即ち入力文書と検索対象文書との間に少しでも類似度を持つならば、その大小に無関係に両文書は類似していると判断される。

【００８１】文書類似度計算部１０は、入力文保存部８に保存された文書と文書記憶部１１から取り出された検索対象文書とが類似していると判断した場合、その旨を制御部２に通知する。すると制御部２は、文書類似度計算部１０により入力文書に類似していると判断された検索対象文書を表示部３に検索結果として表示する（ステップＳ１８）。

【００８２】次に文書類似度計算部１０は、文書記憶部１１内に未処理の文書、即ちステップＳ７で未だ取り出されていない文書が残っているか否かを調べる（ステップＳ１９）。

【００８３】もし、未処理の文書があるならば、文書類似度計算部１０はステップＳ７に戻り、文書記憶部１１から次の１文書を検索対象文書として取り出し、前記したステップＳ８以降の処理を行う。

【００８４】これに対し、未処理の文書がないならば、文書類似度計算部１０は制御部２に制御を戻して処理を

終了する。なお、本発明は前記した実施の形態に限定されるものではない。

【００８５】即ち、前記実施形態では、入力文保存部８に保存された入力文書の各文の重み付けを行うのに、単語の重要度に従う重み付けの手法を適用した場合と、入力文書中で使用されている単語の頻度（使用頻度）に従う重み付けの手法を適用した場合について説明したが、これに限るものではない。例えば、単語の重要度をすべて同じ値（例えば１）に設定すれば、重み付けしない場合と等価であると見なせる。また、単語の重要度と単語の頻度とを組み合わせて重み付けするようにしても構わない。具体的には、入力文書に現れる単語の重要度と当該単語の頻度との積を各文で加算し、その加算結果を該当する文の重みとするようにしてもよい。この場合、頻度１の単語も対象とするとよい。

【００８６】また、前記実施形態では、入力文保存部８に保存される（単語分割、文節合成後の）入力文の群からなる文書、即ち検索キーとなる文書が、入力部１から入力されたものである場合について説明したが、例えば文書記憶部１１に記憶されている複数の文書の中からユーザにより指定された文書であっても構わない。

【００８７】また、前記実施形態では、文書記憶部１１に記憶されている複数の文書のすべてが検索の範囲（検索対象文書）となる場合について説明したが、ユーザにより指定された範囲の文書だけを検索対象文書とすることも可能である。要するに本発明は、その要旨を逸脱しない範囲で種々変形して実施することができる。

【００８８】

【発明の効果】以上詳述したように本発明によれば、指定文書（ユーザの入力した文書あるいはユーザの指定した文書）中のすべての文を検索入力文として、検索対象として用意された文書中の各文に対して、１文全体の類似度を求めて順次加算した値を文書全体の類似度とし、指定文書と検索対象文書とが文書全体としてどれだけ類似しているかを、当該文書全体の類似度で表すことにより、この文書全体の類似度をもとに検索対象となる複数の文書の中から指定文書と類似した文書を検索することができ、ユーザが望む文書と類似する文書を検索する際に、ユーザは検索キーとして単語やキーワードなどを選び出すというような手間が一切必要なくなる。

【００８９】また本発明によれば、指定文書で使用されている各単語の重要度または使用頻度の少なくとも一方に応じて当該指定文書中の各文に重み付けを行い、その各文ごとの重みをその文（検索対象文）と検索入力文との文全体の類似度に反映させることで、すべての検索対象文書の中から指定文書と類似した文書であって且つ指定文書中で使用している重要度の高い単語を多数含む、あるいは指定文書中で使用している単語の使用数の多い文書を検索することができる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る文書検索装置の全体構成を示すブロック図。

【図２】同実施形態の動作を説明するためのフローチャートの一部を示す図。

【図３】同実施形態の動作を説明するためのフローチャートの残りを示す図。

【図４】図３中のステップ６の処理（入力文の重み付け処理）の詳細を説明するためのフローチャートであり、図４（ａ）は単語の重要度により文に重み付けをする場合のフローチャート、図４（ｂ）は入力文書内での単語の使用頻度により文に重み付けをする場合のフローチャート。

【図５】入力部１から入力された検索のキーとなる入力文書を構成する複数の入力文の例と、文書記憶部１１から取り出された検索対象文書を構成する複数の文（検索対象文）の例を示す図。

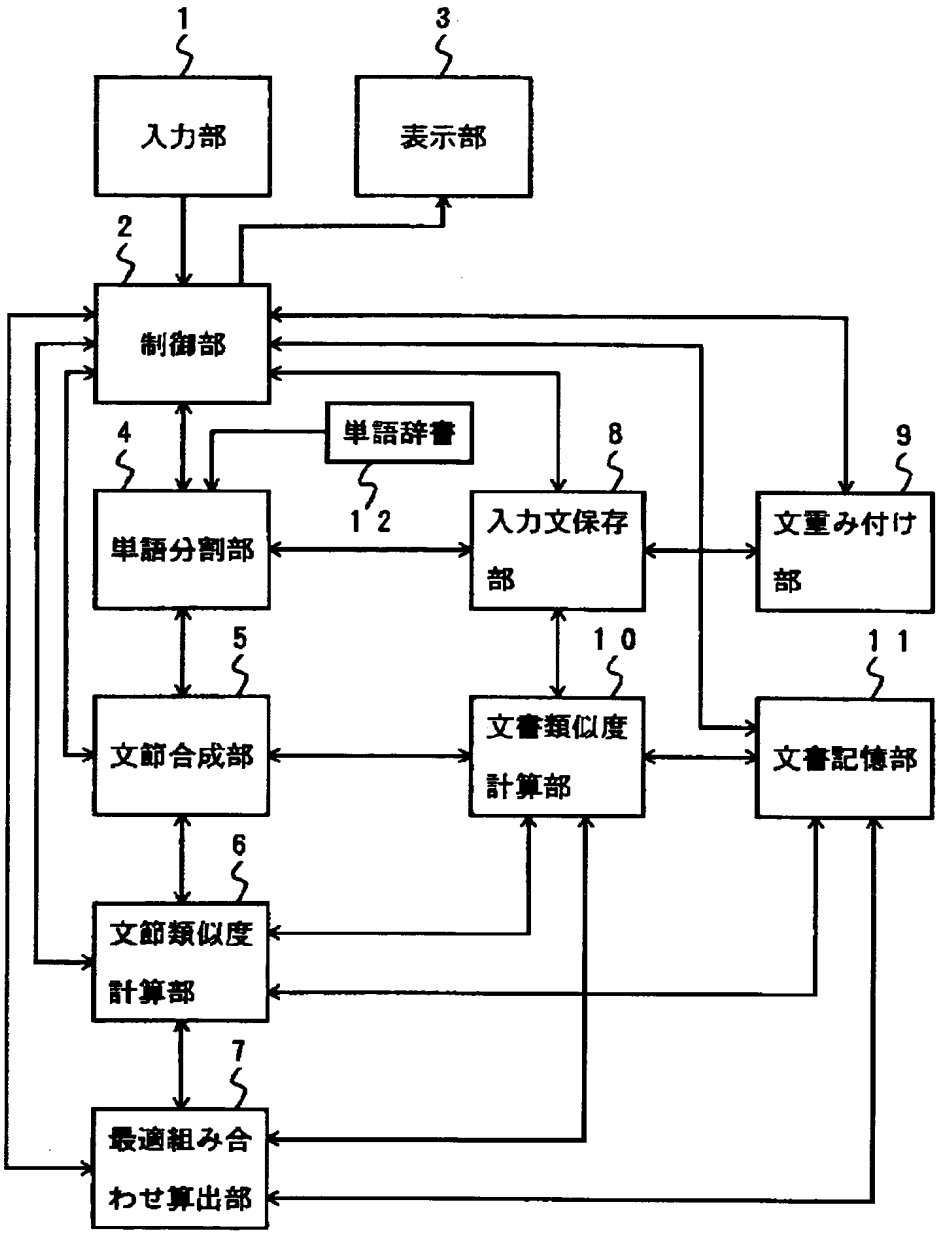
【図６】図１中の単語分割部４による単語分割結果（単語分割された文、当該文の各単語、及び重要度を含む付加情報）の例と、この分割された各単語が文節合成部５により合成されて文節が生成された例を示す図。

【図７】図５中の入力文書の各入力文ａ１～ａ３に文重み付け部９が単語の重要度により重み付けした場合の各入力文ａ１～ａ３の重みの例と、文書類似度計算部１０により求められる図５中の入力文書の各入力文ａ１～ａ３と検索対象文ｂ１，ｂ２との間の文全体の各類似度の例を示す図。

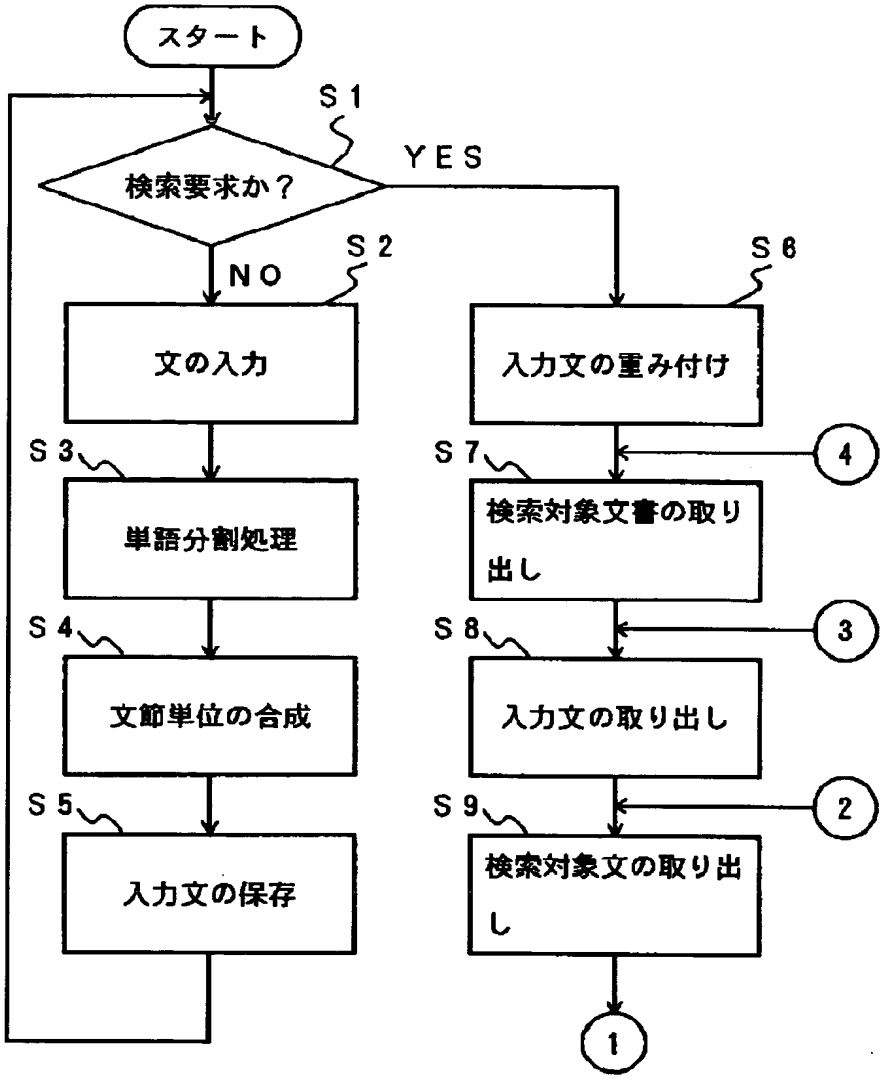
【符号の説明】

１…入力部、
２…制御部、
３…表示部、
４…単語分割部、
５…文節合成部、
６…文節類似度計算部（文構成単位類似度計算手段）、
７…最適組み合わせ算出部、
８…入力文保存部、
９…文重み付け部、
１０…文書類似度計算部、
１１…文書記憶部、
１２…単語辞書、
ａ１～ａ３…文（検索入力文）、
ｂ１，ｂ２…文（検索対象文）。

【図１】


1
入力部
3
表示部
2
制御部
4
単語分割部
単語辞書
12
8
入力文保存部
9
文重み付け部
5
文節合成部
10
文書類似度計算部
11
文書記憶部
6
文節類似度計算部
7
最適組み合わせ算出部

【図２】



【図７】

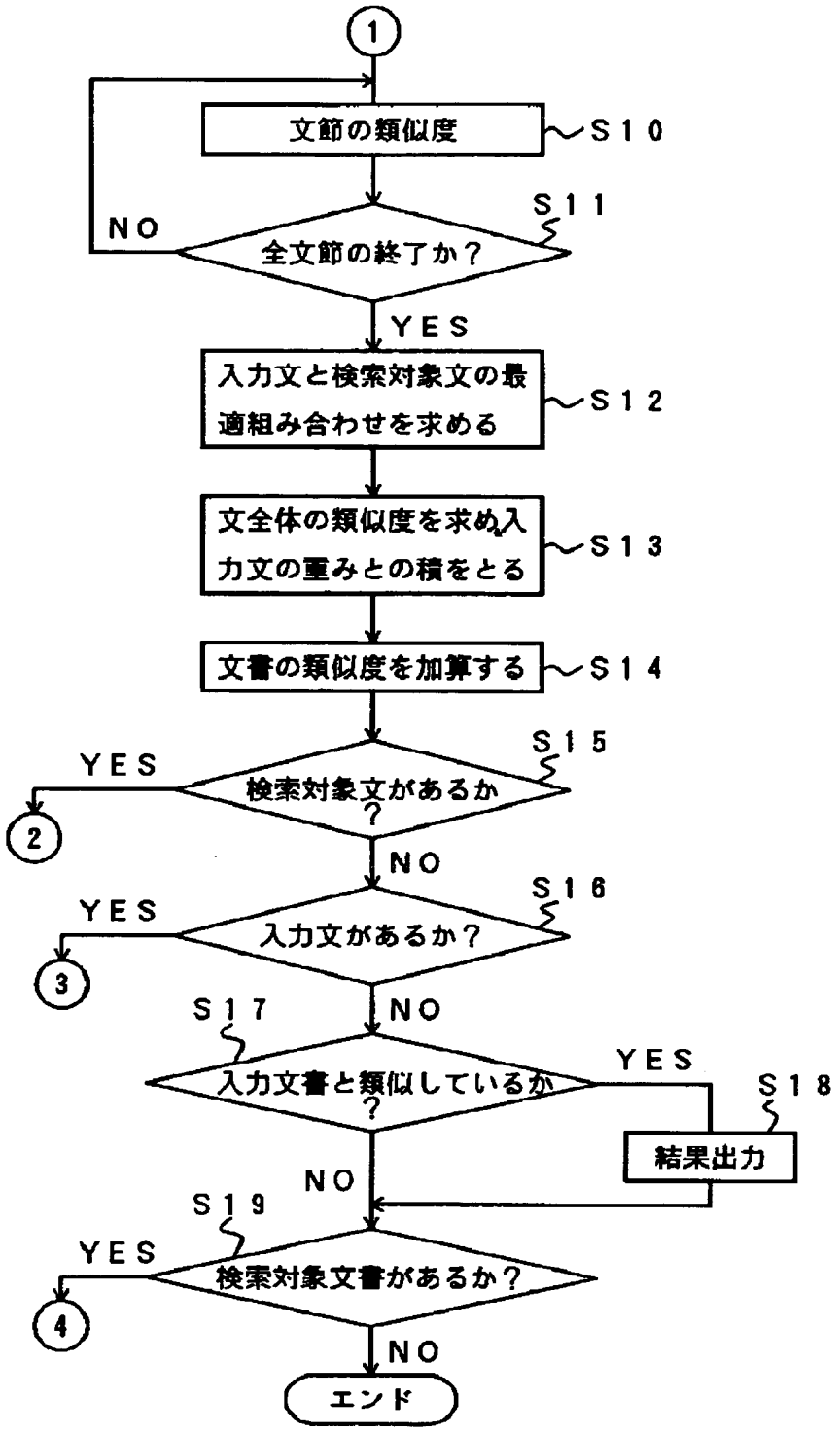
(a)

入力文	重　み
a 1	4
a 2	3
a 3	2

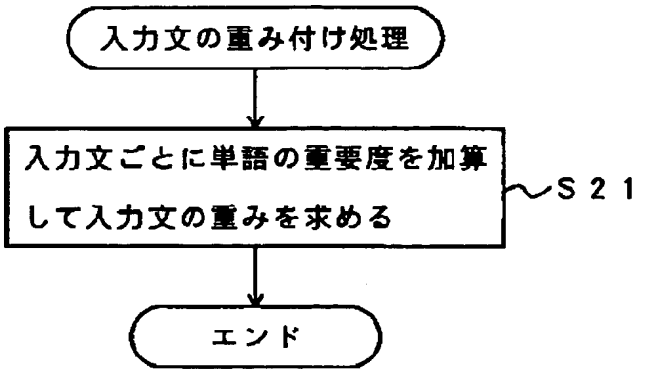
(b)

検索対象文 / 入力文	b 1	b 2
a 1	0.6	0.4
a 2	0.25	0.375
a 3	0.14	0.57

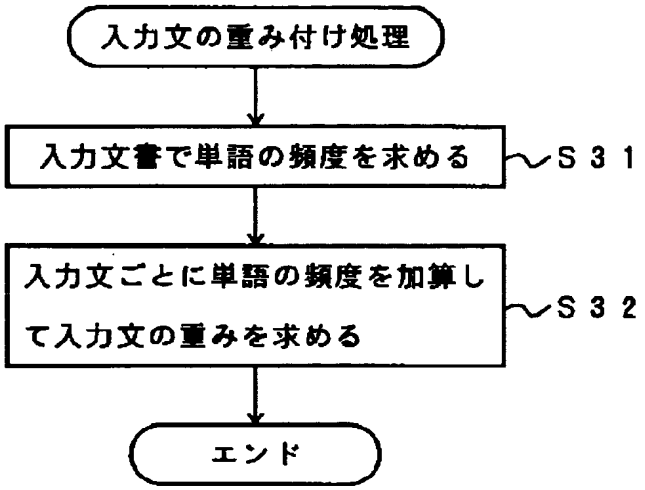
【図３】


1
文節の類似度
S10
S11
NO
全文節の終了か？
YES
入力文と検索対象文の最適組み合わせを求める
S12
文全体の類似度を求め入力文の重みとの積をとる
S13
文書の類似度を加算する
S14
S15
YES
検索対象文があるか？
2
NO
S16
YES
入力文があるか？
3
S17
NO
入力文書と類似しているか？
YES
S18
結果出力
NO
S19
YES
検索対象文書があるか？
4
NO
エンド

【図４】



（ａ）



（ｂ）

【図６】

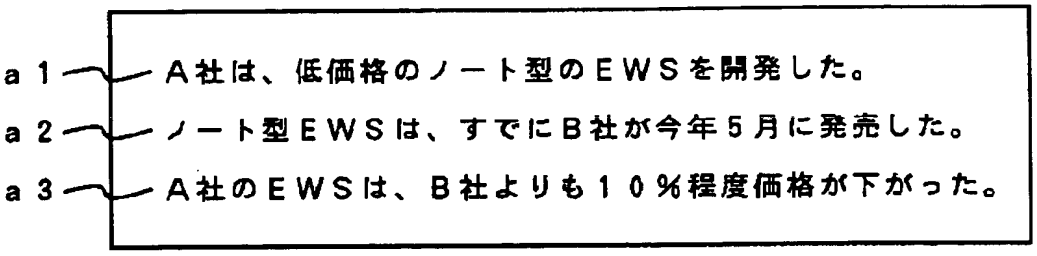
A社は低価格のノート型のＥＷＳを開発した。			
No.	単　語	品　詞	重要度
1	A社	名　詞	1
2	は	助　詞	0
3	低価格	形容動詞	0
4	の	助　詞	0
5	ノート型	名　詞	2
6	の	助　詞	0
7	EWS	名　詞	1
8	を	助　詞	0
9	開発する	動　詞	0

（ａ）

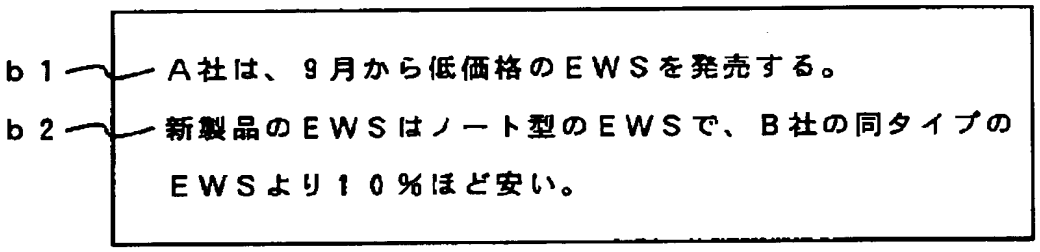
1	A社は
2	低価格の
3	ノート型の
4	EWSを
5	開発する

（ｂ）

【図５】


a 1
Ａ社は、低価格のノート型のＥＷＳを開発した。
a 2
ノート型ＥＷＳは、すでにＢ社が今年５月に発売した。
a 3
Ａ社のＥＷＳは、Ｂ社よりも１０％程度価格が下がった。

（ａ）


b 1
Ａ社は、９月から低価格のＥＷＳを発売する。
b 2
新製品のＥＷＳはノート型のＥＷＳで、Ｂ社の同タイプの
ＥＷＳより１０％ほど安い。

（ｂ）